BRINGING PATIENT POPULATIONS TO THE INTEGROME

Isaac Kohane[1,3,4], Shawn Murphy[1,4], Atul Butte[2], Vlad Vlatchinov[1]

1. Partners Healthcare System; 2. Stanford School of Medicine; 3. i2b2 NCBC; 4. Harvard Medical School Center for Biomedical Informatics

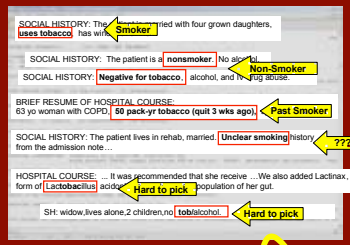
ABSTRACT

The Integrome project of i2b2 has as its goal the redefinition of human disease based on combined genomic and clinical data. Why is this important? The current classification of disease is a mix of attributing specific diseases to specific organ systems (e.g. heart disease) or to specific clinical manifestations (e.g., diabetes mellitus, named after the presence of glucose in the urine). These classifications have been the mainstay of medical education and practice for decades and even centuries. As a result, the commonalities across these diseases, the underlying pathophysiological processes that span these various diseases are only glimpsed at from the particular perspective of this historically accreted rather arbitrary classification scheme.

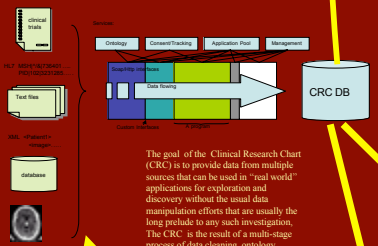
As an initial foray into developing a data-driven robust view of disease that includes all genomic data and clinical findings, *the Integrome*, i2b2 investigators obtained all the measures of genes in the National Library of Medicine's Gene Expression Omnibus, a public database and used artificial intelligence techniques to read the textual descriptions of these tens of thousands of experiments and assign these experiments automatically to one or more disease or process categories (e.g., heart failure or aging). In parallel, the investigators took the tens of thousands of genes that were measured in each of these experiments and determined through millions of calculations which genes were truly characteristic of the process or disease described in that experiment and those that were not. This result provided several interesting insights. For example diseases such as gastritis had gene signatures remarkably similar to those of heart attacks, and several genes widely known to be associated with inflammation were associated with a very large array of diseases.

Now, we are integrating genomic measures to patient and population-specific categories. The clinical phenotypes of the 2.5 million patients in the Partners Health Care system's electronic medical record (extracted by Natural Language Processing from clinical notes and from structured fields such as clinical labs, with due protection of patient privacy) are being uploaded into the i2b2 Clinical Research Chart. This next step will not only allow us to understand how we can reclassify diseases based on their genomic signatures but will allow us to determine how individual patients can be recategorized based on their relationship to one another within the Integrome, thereby directly informing diagnostic and prognostic opportunities. Importantly, this Integrome will also identify therapeutic modalities that have broader scope due to shared molecular motifs across pathophysiology. In doing so several important challenges will have been overcome that all similar investigations must address:

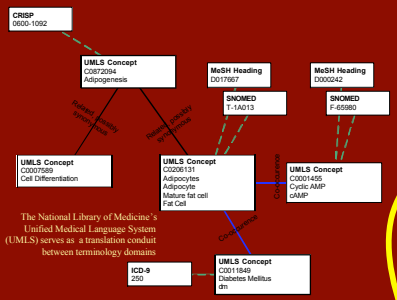
• Consent granularity and persistence	• Noise-robust probabilistic discovery
• Extracting crisp phenotypes from imprecise textual data.	• Wrapping complex analyses within simple-to-use workflows for investigators.
• Integration across multiple conventional clinical and genomic data types	• Structure-function inference for disease prediction from genomic and protein variants.
• Consistent & durable collaborations between experts in information science and biomedical research.	• Making translational technologies available to underserved populations.



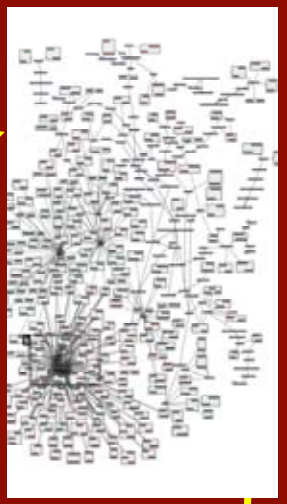
A large fraction of clinical data in electronic form is present as unstructured text fields. Consequently, i2b2 investigators have invested significant resources in developing Natural Language Processing (NLP) tools that can extract robust phenotypes from clinical records to then be entered into the Clinical Research Chart database. These NLP tools are made available as Hive services.



The goal of the Clinical Research Chart (CRC) is to provide data from multiple sources that can be used in "real world" applications for exploration and discovery without the usual data manipulation efforts that are usually the long prelude to any such investigation, The CRC is the result of a multi-stage process of data cleaning, ontology matching, consent tracking, anonymization and multi-data source integration with annotation of data provenance and transformation.



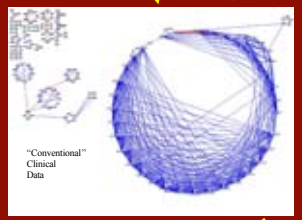
The National Library of Medicine's Unified Medical Language System (UMLS) serves as a translation conduit between terminology domains

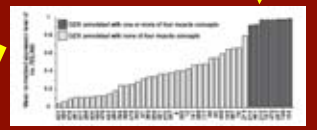


A taxonomy of Pathophysiology firmly grounded in the molecular signatures that diseases and biological processes share and those signatures that distinguish them.

The i2b2 "Hive" provides Core Computational services for clinical research including population selection functions, analytic functions, support for the Clinical Research Chart, and overall workflow sequencing. Many of the Services delivered by the Hive are programs written by other groups "wrapped" in a Hive protocol

Deconstructing the NCBI's Gene Expression Omnibus into Gene expression measures and disease/biological process groups

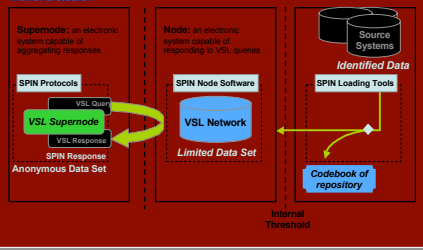




Gene product biomarkers understood as a continuum rather than thresholded discrete variable

Shared Pathology Informatics Network and Virtual Specimen Locator

Data Flow and Tool Sets: HIPAA compliance via programming, data location, and firewall protection



POPULATION-WIDE & POPULATION-SPECIFIC INTEGROME

- Biomarkers (and combinations thereof) that are maximally sensitive with reduced false positive rate
- Convenience samples of very large size stratified by clinical and genomic characteristics.
- Probabilistically-ranked mono- & multi- genic disease/biological process hypotheses.
- Availability of large biosample data sets for validation.
- Identification of common therapeutic opportunities across previously distinct disease classes.